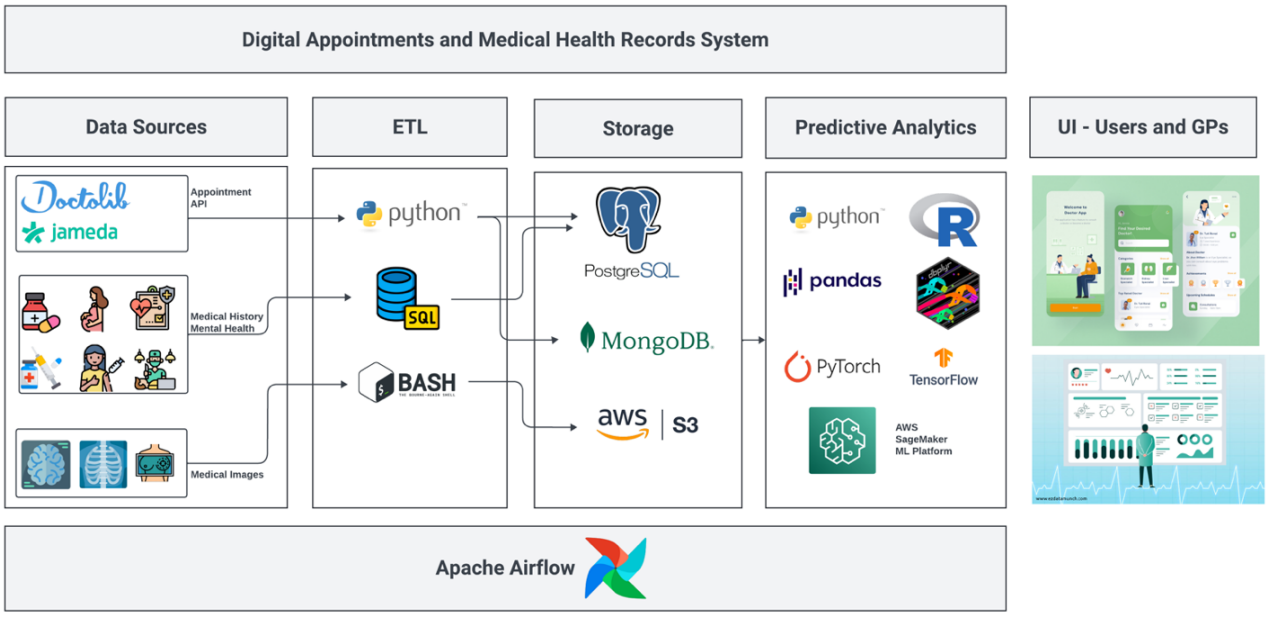
Digital Appointments and Medical Health Records System



Date: 04.12.2023
Course: Big Data Challenge

Submitted by:

Sifael Sebastian Ndandala
Pakin Veerachanchai
Sripathy Kathiresan Tamilselvan
Nishant Parmar

Introduction

Healthcare is a vital part of our infrastructure, contributing directly to the overall well-being and quality of our day-to-day lives. In Germany, while good quality healthcare services are available, access to general practitioners or specialized doctor visits can be painfully difficult as appointment openings are often several weeks out. This can often mean not being able to see a doctor when you need them or have an appointment long after recovery.

Furthermore, healthcare data such as medical history, medical images, and mental health records remain partially digitized and fragmented, making it difficult for practitioners to wholistically assess the patient's complete medical history and provide individualized care. In addition, the lack of digitization is a lost opportunity to use modern technologies in machine learning and deep learning which are becoming increasingly effective in producing accurate predictive diagnosis. Digitizing personal health records can therefore improve individualized medicine as new technologies provide more accurate diagnosis and reduce misdiagnosis and other medical errors.

Our goal with the Digital Appointments and Medical Health Records System is to tackle these challenges thereby improving the experience of patients and general practitioners.

1. First, we develop a digital appointment system that allows patients to access practitioners and doctors in their local area by combining availabilities from known providers such as DoctorLib and Jameda, all in one app.

2. Second, adding to the digital appointment, we digitize medical and health records of patients within the app allowing them to know and have better discussions with their service provider. Furthermore, patients using the appointment system can give access to their practitioner ahead of their appointments, all within the app, making appointments more effective.

The Digital Appointments and Medical Health Records System benefits both users/patients by giving them access to their medical data and better availability of local practitioners. For practitioners and specialist, they can now have more access to new patients based on their availability, and most importantly, they can request access to medical history ahead of a visit and leverage predictive technologies to enhance treatment and their services.

Project Technical Scope

The technical scope is primarily driven by the data we need to successfully develop and integrate appointment scheduling and medical health records into one workflow. Therefore, we limit our scope to two core areas:

1. **Digital Appointment Scheduling**

To achieve better appointment availability and listing, we use API to known appointment providers, **Doctorlib** and **Jameda**, thereby allowing us to compare against the platforms and offer better appointment openings to prospective patients.

2. **Medical Health Records**

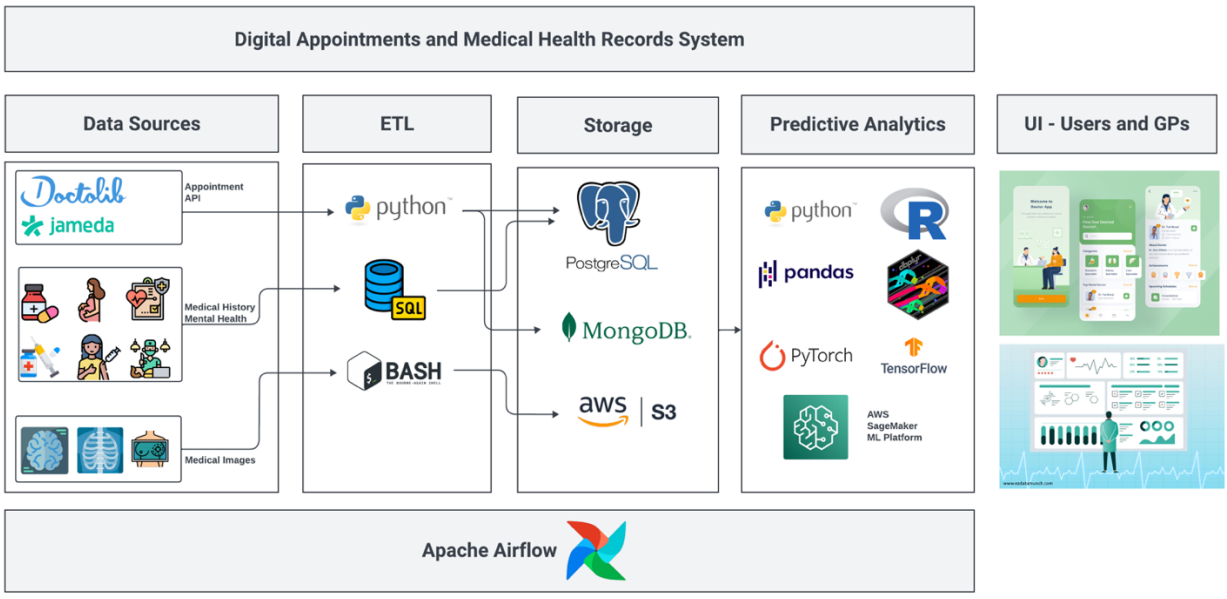
Medical health records can be vast. For this project, we limit the medical data into three categories: medical imaging, mental health data, and overall medical history. These categories offer the best combined use of predictive analytics particularly in determining diagnosis.

Data Pipeline Operational Workflow

Our data pipeline is implemented in four key stages namely:

- **Data Sources**: Sources of our data
- **ETL**: Extraction, Transformation, and Loading of Data
- **Storage**: Technologies for storing and retrieving data
- **Predictive Analytics**: User of Machine Learning for Predictions

The workflow chart below outlines each stage and the respective technologies. In the following section, we discuss each stage and offer further technical details.



1. Data Sources

Our main objective is two-fold: digital appointments system and medical health records. Consequently, our data sources and collection are grouped by these functions.

1.1. Digital Appointment System

Our goal with the appointment system is to offer users appointment options with local general practitioners with an emphasis on experience and the nearest dates. To that end we use two important data sources.

a) API Integration with DoctorLib and Jameda

DoctorLib and Jameda are established platforms that offer scheduling services with general practitioners throughout Germany. Through API integration, we connect our system with their databases, gaining access to their combined lists of general practitioners and their appointment openings. Specifically, we collect the following data points:

- **Availability:** Up-to-date appointment availability of the practitioner
- **Location:** Current address and/or zone of operation
- **Specialty:** The specialty in which the GP/Doctor practices.

b) Doctor Rating

Patient reviews can be valuable for future patients when deciding on their medical services. Jameda provides practitioner reviews, and we collect this data to add to the appointment system.

1.2. Medical Health Records

To consolidate a patient's medical history and records, we gather medical data from various sources including clinics, labs, personal reports, and insurance companies. For technical pipeline reasons, we categorize them by their data types: medical imaging, medical history, and mental health.

a) Medical Imaging

Medical imaging includes **MRI, X-Ray** and **CT scans**. These elements will provide a wide range of rich imaging data to evaluate patient's history to inform best care for the patient. Similarly, they offer opportunity to leverage machine learning and deep learning to assist with accuracy in diagnosing a patient.

b) Mental Health

Mental health can have far-reaching impact on physical wellness. As part of data collection, we consider **Therapy Sessions** and **Mental health-related medication** such as Anti-depressants to add to the medical record.

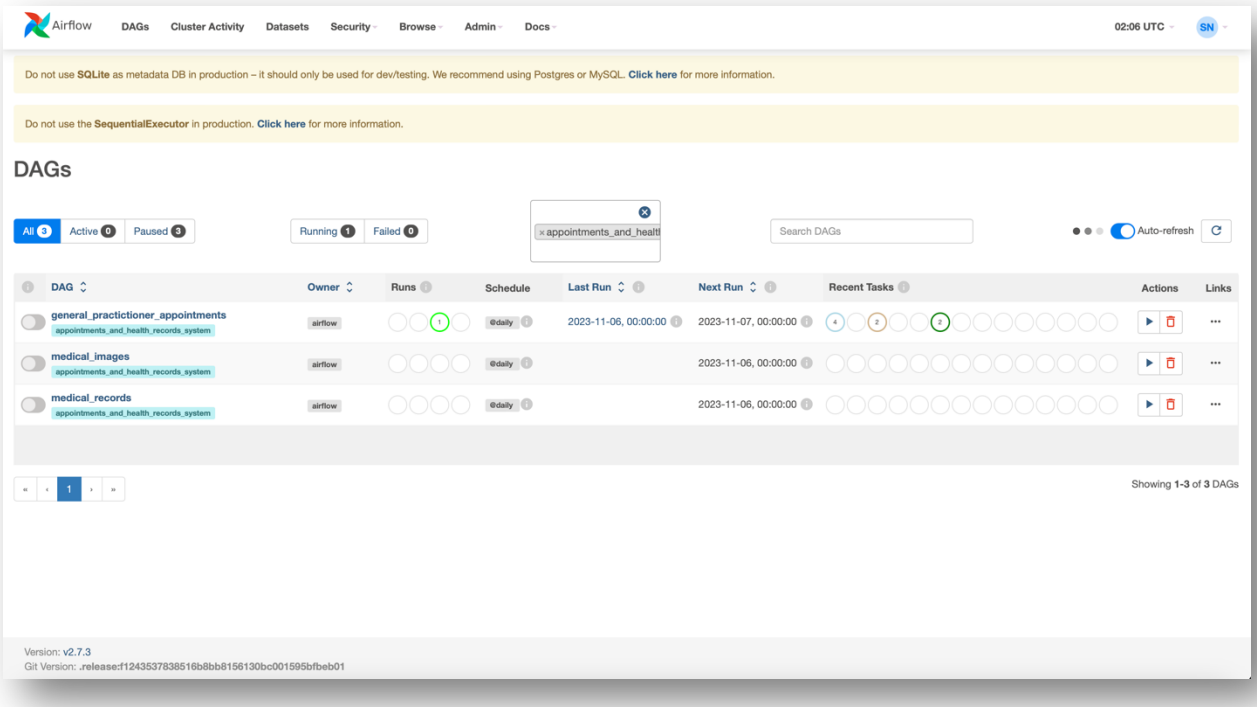
c) Medical History

The medical history data will include the following:

- **Diagnoses**: The app will record and organize information on past diagnoses, aiding healthcare providers in making informed decisions based on a patient's medical history.
- **Medical Tests and Blood Samples**: Tracking blood sample data helps to monitor certain conditions and provides a thorough picture of a patient's general health.
- **Vaccinations**: The app will keep track of immunizations, providing regular updates and information to support preventive care.
- **Allergies to Medication**: By recording allergies, the app makes sure doctors only prescribe appropriate and safe drugs.
- **Major Surgeries or Medical Events (Childbirth)**: A comprehensive record of significant medical procedures and events, such as births, gives insight into a patient's entire medical trajectory.
- **Current Medication:** By keeping track of the drugs, a patient is currently taking, the app will encourage adherence to treatment and help to avoid drug interactions.

2. ETL: Data Pipeline and Processing

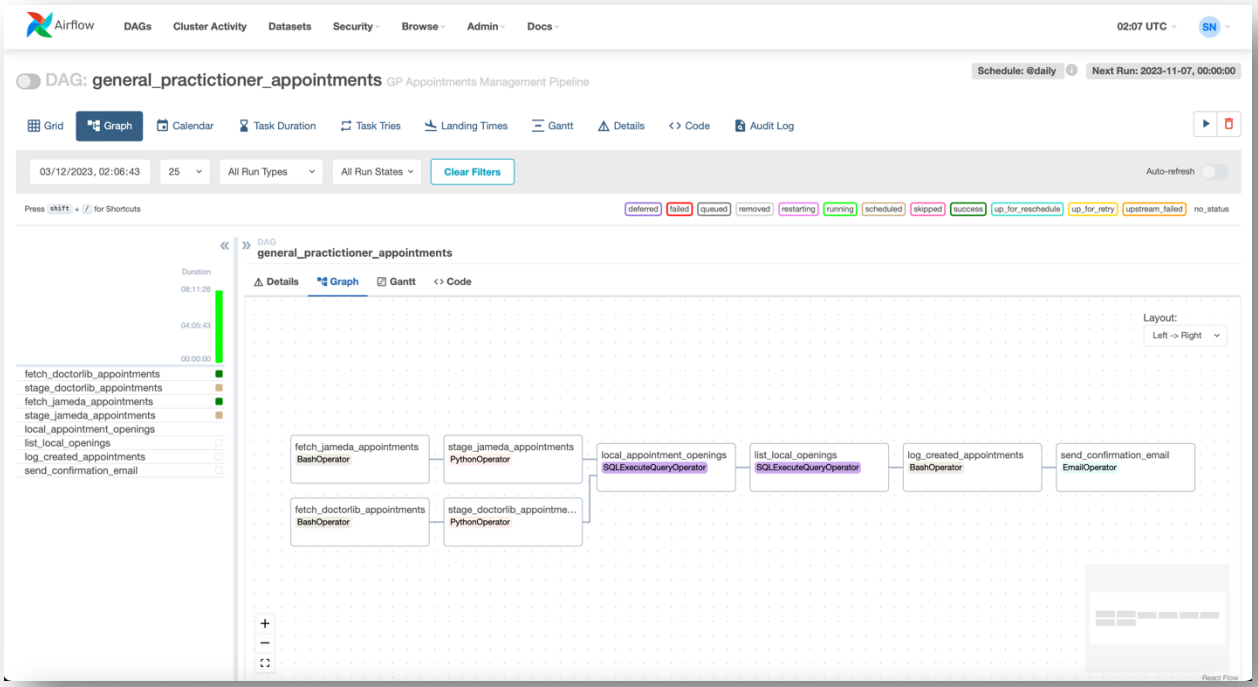
To handle the complexity of data types and variety of data sources, we developed three pipelines. These pipelines cover the appointment system, medical history and medical images respectively, as depicted below and demonstrated in the airflow DAGS.



Pipeline DAG 1: General Practitioner Appointments

The appointments pipeline uses APIs from **Doctorlib** and **Jameda** to retrieve information about general practitioners, their availability, location, and ratings. The pipeline consists of the following tasks:

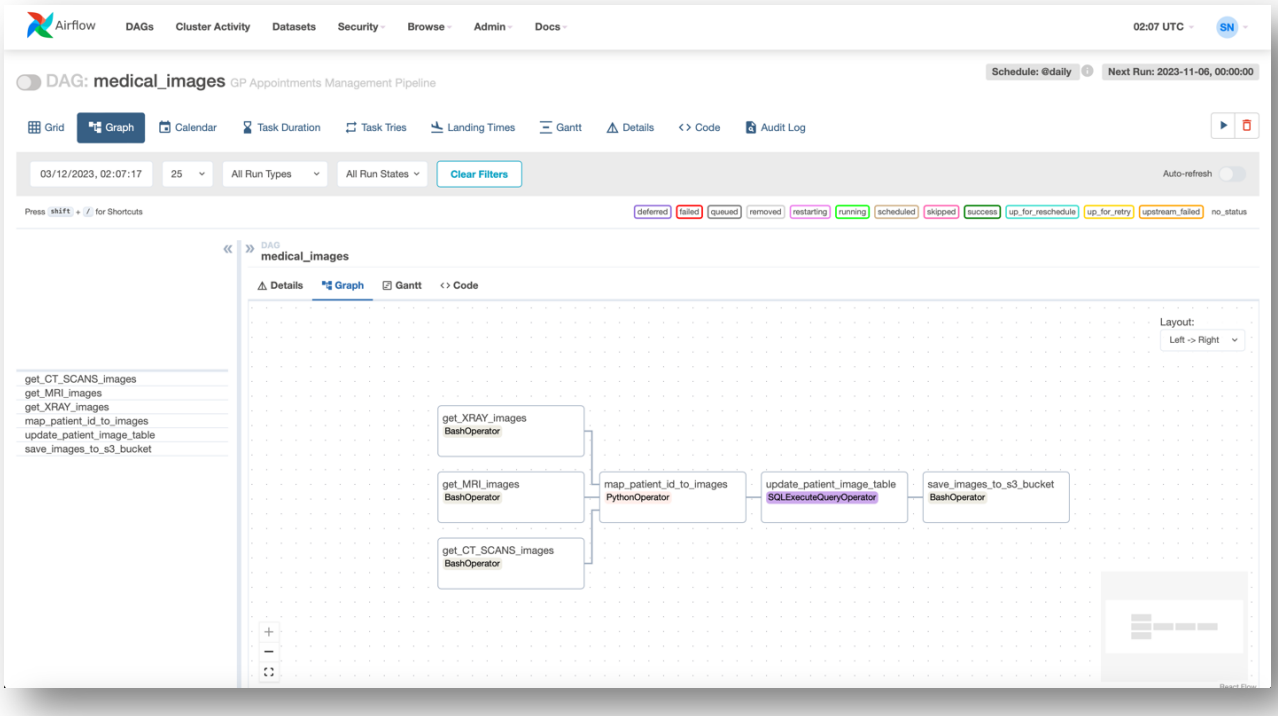
1. **API Data Retrieval**: The initial two tasks that retrieve data from the Doctorlib and Jameda API to a local file system for processing. For simplicity, we use BashOperator to download files in json format.
2. **Staging Tasks**: As the two platforms may return varying structure and results, each result is passed to a **staging task** for cleaning and normalization before merging the data into a single SQL table. PythonOperator is used for this staging step.
3. **local_appointment_opening**: This task combines normalized data from Doctorlib and Jameda into a single SQL table with all local appointment openings.
4. **list_local_openings**: This task runs a query to list local appointments available to display to the user.
5. **log_created_appointments**: When a user creates an appointment, this task logs the appointment details into an appointments table and cancels that opening.
6. **send_confirmation_email:** This task uses an EmailOperator to send a confirmation email to the user, detailing their appointment time, location, and date.
7. **share_medical_history:** This task displays to the user if they would like to share their medical data with the general practitioner after setting up their appointment.



Pipeline DAG 2: Medical Imaging

Medical Images require separate treatment from other data types, as images need to be stored independently of the databases. To manage image processing, our pipeline downloads the images into a staging folder, where a mapping of the images and their respective patient IDs is performed. This mapping is processed in Python and stored in a SQL table to join with non-imaging medical health data later. The pipeline task are as follows:

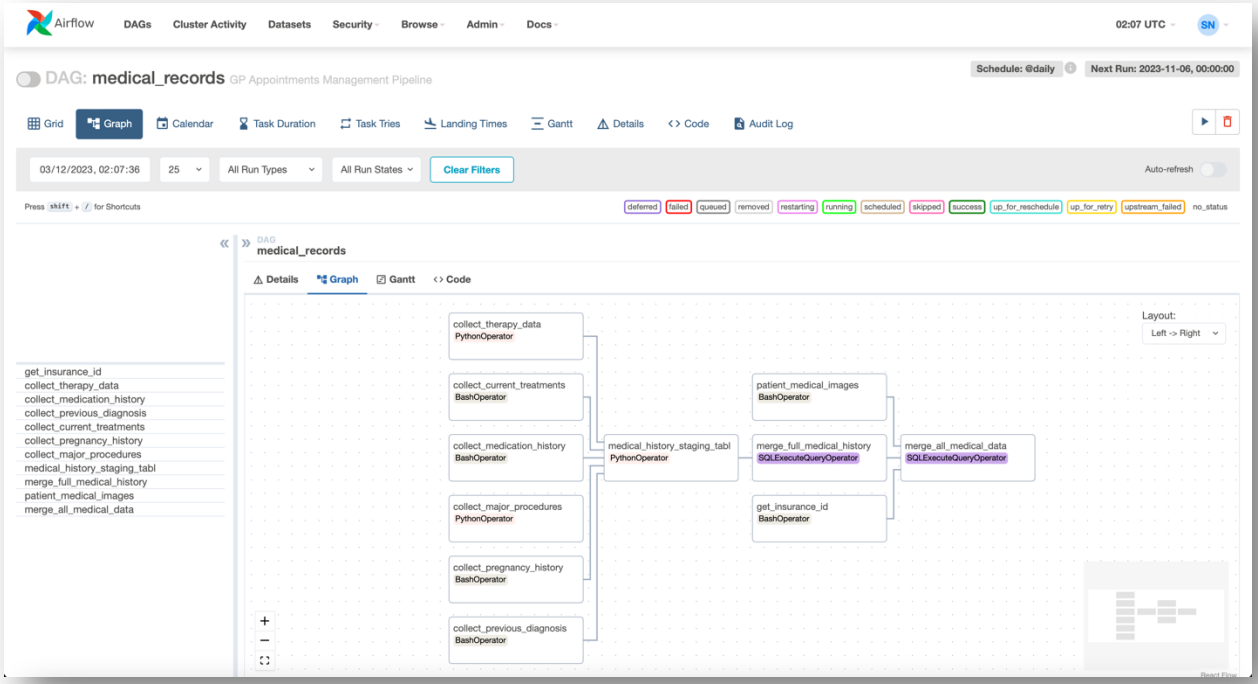
1. **Data Retrieval**: These are three tasks that retrieve medical image data. Using bash scripts, they download x-ray, MRI and CT scans into a local folder for processing.
2. **map_patient_id_to_images**: This task uses python to create a map between all image files and the patient id. This information will be used to join with non-medical health data.
3. **update_patient_image_table:** This task takes the mapping from the previous tasks and stores/updates it in a SQL table.
4. **save_images_to_s3_bucket:** Images cannot be saved in tables; therefore, it is necessary to store them in an easily accessible solution. In our case we use, Amazon's S3 bucket.



Pipeline DAG 3: Medical History

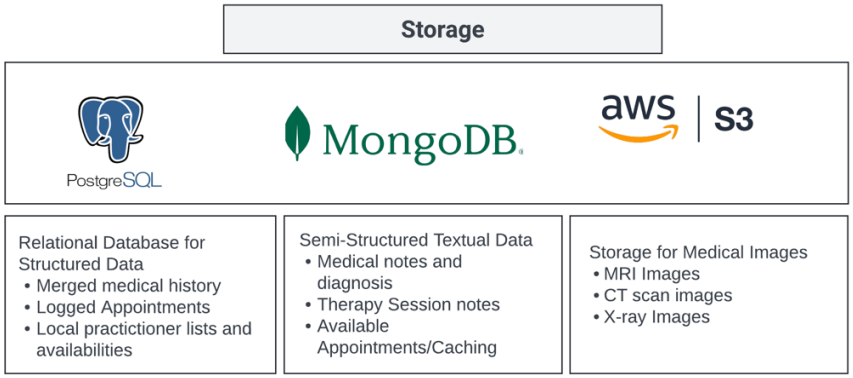
Medical history covers a variety of text-based data and therefore requires a large staging process to combine all medical records under their patient ID. In the pipeline, each individual task retrieves specific medical record (vaccination, major medical procedures) and stores them in a staging folder for cleaning and mapping to patient ID. Once the mapping is complete, the medical history is joined together into a SQL database for on-going storage.

A further step in this pipeline merges the medical imaging data (through image ids) from pipeline #2 and insurance id to complete the mapping of the user's complete medical data. The implementation of the pipeline is given below.



3. Data Storage

The project's complexity and variety of the data requires the use of three separate solutions for various data types and needs. For this project we used three technologies: **PostgreSQL**, **MongoDB** and **AWS S3**.



PostgreSQL

PostgreSQL is our primary solution for all medical and health data. The schema of the medical records is based on a patient id as the primary key making it possible to search and retrieve health information including medical image ids that can be joined with other solutions to serve images on the application.

MongoDB

MongoDB is a fast and effective storage solution for semi-structured datasets that may require a high degree of caching and streaming feeds. This makes this an ideal solution for managing the appointments system as it is highly scalable and can be used for real-time application. We also leverage MongoDB for document-based storage including medical notes, therapy session notes and prescription.

AWS | S3

Unlike text and documents, images cannot be stored in a traditional database. Nonetheless, we need a way to quickly access images served as static files to our user interface. AWS S3 is a solution for storing files in folder structures that can also be retrieved easily as static files for rendering and visualization.

4. Predictive Analytics

The motivation to digitize health data does not end with being able to share and discuss it with a general practitioner. In modern times, it is just the beginning. With deep learning and AI permeating through all sectors, predictive analytics can play a vital role in improving personalized healthcare.

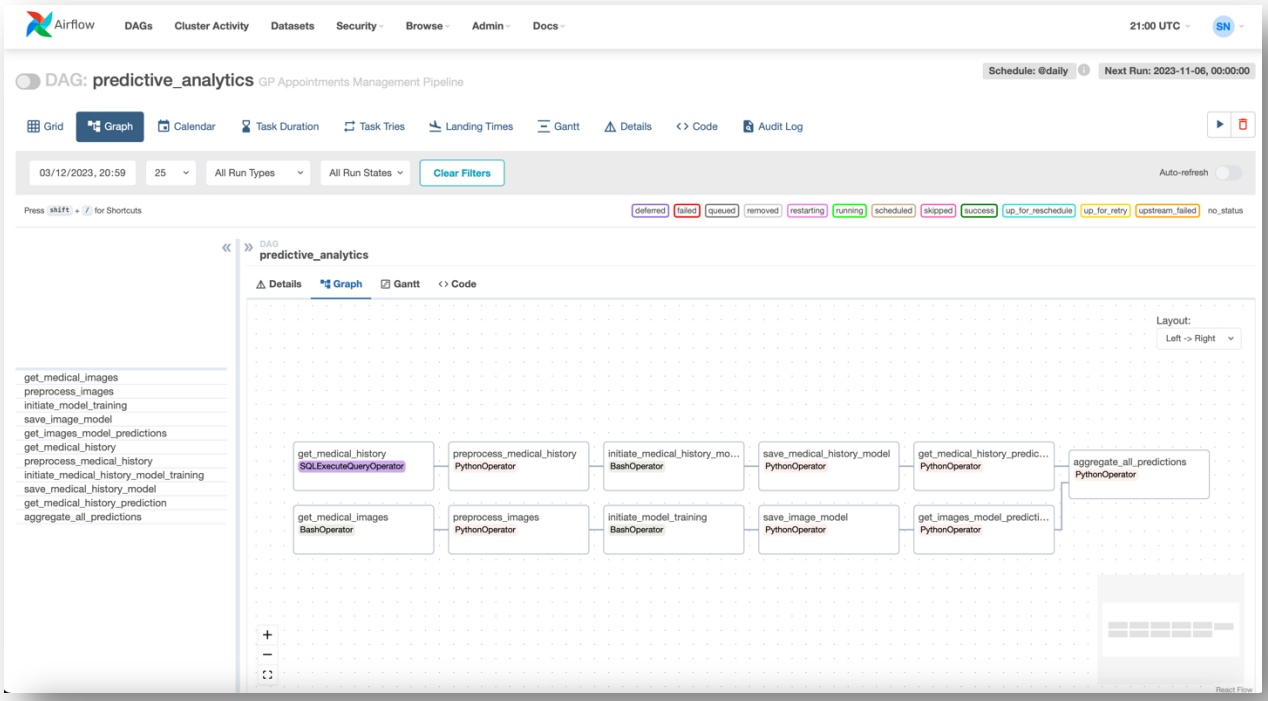
While deep learning and AI can look at several outcomes, the scope for our project is to use the medical history and medical images to predict ailments of the patient and provide practitioners with tools to improve diagnosis, catch health problems early on, and reduce errors.

The predictive analytics pipeline trains predictive models using medical images and health records separately and generates predictions respectively. Once predictions of diagnosis are generated, the pipeline combines the predictions into an easily navigable user interface for general practitioner use.

The pipeline tasks are described below:

1. **Fetching Data (get_medical_history** and **get_medical_images):** These tasks retrieve medical history from earlier pipelines and medical images stored in S3 buckets. We use SQLOperator and BashOperator respectively.
2. **Preprocessing (preprocess_medical_history** and **preprocess_images):** These tasks implement machine learning preprocessing steps ahead of model training. This work can be executed in python or R.
3. **Model Training:** The model training tasks **initiate_image_model_training** and **initiate_medical_history_model_training** are used to initiate model training and assess deep learning models. The training happens on the cloud. We use Pytorch or Tensorflow for training.

4. **Save Model**: Model performance is evaluated at the model training and the best model is selected to run predictions.
5. **Generating Predictions**: Each saved model then generates predictions of the new patient both from images and history. Predictions can be multiple diagnosis.
6. **Aggregate Predictions**: Prior to this step, each model has been running independently. At this step, all predictions are aggregated and ready to be displayed in the doctor's UI for evaluation, official diagnosis and discussion. We save the predictions on a MongoDB database to allow for real serving of the predicted outcomes.



Risks, Challenges and Considerations

Healthcare is a highly regulated sector and for good reason. Inadequate controls can lead to significant impact on individual lives. Therefore, digitization of medical and health records and appointment systems need to exam risks, challenges and considerations to enhance healthcare services and limit harm. Below is a discussion on associated challenges and risks with our digitization project.

Digital Literacy and Accessibility:

A significant difficulty to digital health implementation is lack of digital literacy among users. The adequacy of digital health literacy influences not just the ability to actively engage with digital services, but also other factors such as self-care, being secure and in control, motivation to participate with

digital services, and access to effective systems. (Kemp et al. 2020) The development of health applications must emphasize some users lacking digital health literacy, as well as to provide alternatives for users with limited digital health literacy.

Also, with lower levels of digital health knowledge, older age is often considered a barrier to digital health use. However, this is not always the case. This issue is also influenced by consumer preferences and incentive to engage in digital health approaches (Kemp et al. 2020). Increased digital health knowledge in elderly people will enable successful digital health deployment. Not to mention, understanding customer's choices and not assuming they are unable or uninterested in digital health technologies are primary goals for digital health literacy.

The implementation of digital health should consider the wide range of digital health literacy, particularly among disadvantaged individuals as well as all ages populations. To enhance access and accessibility, it is important to execute design and cooperation strategies, and health policy should align with this strategic approach.

Data Security and Privacy:

Healthcare providers need to safeguard huge amounts of data for efficient and effective patient care. Data breaches, especially in healthcare, pose serious risk as attackers dig up and release sensitive data. Security mechanisms including Authentication, Encryption, and Data masking are becoming more effective (ABOUELMEHDI et al. 2017), yet advanced attacks still target the healthcare sector. As digitized healthcare system, enterprises need sophisticated healthcare data security solutions to protect valuable assets and comply with regulations.

Advanced threats threaten data security, especially in big data analytics. Data security controls access throughout the data lifecycle, while data privacy follows policies and legislation to restrict personal data access. Healthcare providers must follow varied data protection rules to protect personal data legally.

Balancing Convenience with Privacy:

Finding the right balance between system usability and privacy is important in the realm of digital healthcare. While ensuring that users can easily navigate and utilize the system is important for its effectiveness, it is equally vital to prioritize the protection of sensitive patient information. Minimizing personal data requests is essential for addressing digital healthcare privacy challenges in the susceptible wireless context. Additionally, the technique is effective for those in good health who may be less reluctant to provide personal information. It addresses the personalization-privacy problem by automatically selecting the best features (Lee and Kwon 2015).

Limited Data Sharing:

Privacy concerns for healthcare data sharing can be categorized into three groups: defining privacy, implementing privacy regulations, and achieving privacy protection regarding data sharing. These three-layered frameworks demonstrate the ability to clearly define and track private data, as well as through monitoring and data usage limitations (Yang et al. 2012). These frameworks also ensure that

privacy regulations are effectively implemented by establishing guidelines and protocols for data sharing.

The implementation of secure data sharing protocols and encryption methods can be utilized to enhance the exchange of healthcare information across healthcare providers while also protecting the confidentiality of the information. This allows healthcare practices to collaborate and exchange necessary patient information, while ensuring that the security and privacy of the data are maintained. The difficulties in exchanging datasets across healthcare practitioners are obvious. These challenges include data accessibility, interoperability limits, institutional regulations, and technical barriers (Azarm-Daigle et al. 2015). Overcoming these barriers becomes essential for attaining effective healthcare information exchange across authorized providers.

Patient Autonomy:

Patient autonomy in ethical considerations in implementing digital health applications is another important aspect. Ensuring patients control their own health data and can make informed decisions about its use is crucial. This includes obtaining consent for data sharing, providing transparency about how the data will be used, and allowing patients to easily access and manage their own information. The shift towards digitalized healthcare system brings about intricacies and distinctions when compared to conventional healthcare (Schmietow and Marckmann 2019). Individual, collective, patient, and consumer autonomy, digital literacy, democratization, technological reliance, and datafication must all be considered in ethical frameworks.

Conclusion

The Digital Appointments and Medical Health Records system is designed to improve healthcare service for both patients and healthcare providers. By integrating digital appointments and digitizing healthcare data into one app, users can easily assess local appointments and user their data to receive better individualized service. Similarly, practitioners can request access to user data ahead of appointments and leverage machine learning to improve wholistic diagnosis and develop better treatment plans.

To develop this application, we used multiple technologies for storage, data processing and management, covering variety of data sources and data types, all orchestrated in Apache airflow.

The project is cognizant of the risks and challenges with health care regulation, data privacy, digital literacy and data security. In many ways, our implementation has implicit safeguards to privacy and security offering the user ultimate control of their own data. Ultimately, product feedback can help us to improve on the features that best enhance the user and practitioner experience.

References

ABOUELMEHDI, Karim, et al. “Big data security and privacy in healthcare: A Review.” *The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017)*, 2017, p. 8.

Azarm-Daigle, Mana, et al. “A Review of Cross Organizational Healthcare Data Sharing.” *The 5th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2015)*.

JIN, HAO, et al. “A Review of Secure and Privacy-Preserving Medical Data Sharing.” *This work was supported in part by the National Science Foundation of USA under Grant 1547428, Grant 1738965, Grant 1450996, and Grant 1541434.*, 2019, p. 14.

Kemp, Emma, et al. “Health literacy, digital health literacy and the implementation of digital health technologies in cancer care: the need for a strategic approach.” *Health Promotion Journal of Australia*, 2020, p. 11.

Lee, Namyeon, and Ohbyung Kwon. “A privacy-aware feature selection method for solving the personalization–privacy paradox in mobile wellness healthcare services.” *Expert Systems with Applications*, p. 8.

Schmietow, Bettina, and Georg Marckmann. “Mobile health ethics and the expanding role of autonomy.” *Medicine, Health Care and Philosophy (2019) 22:623–630*, 2019.

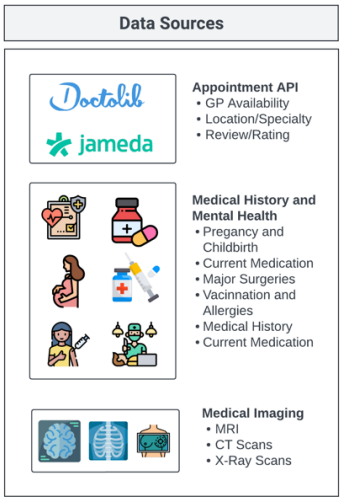
Yang, Ji-Jiang, et al. “A framework for privacy-preserving healthcare data sharing.” *2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)*.

Appendix

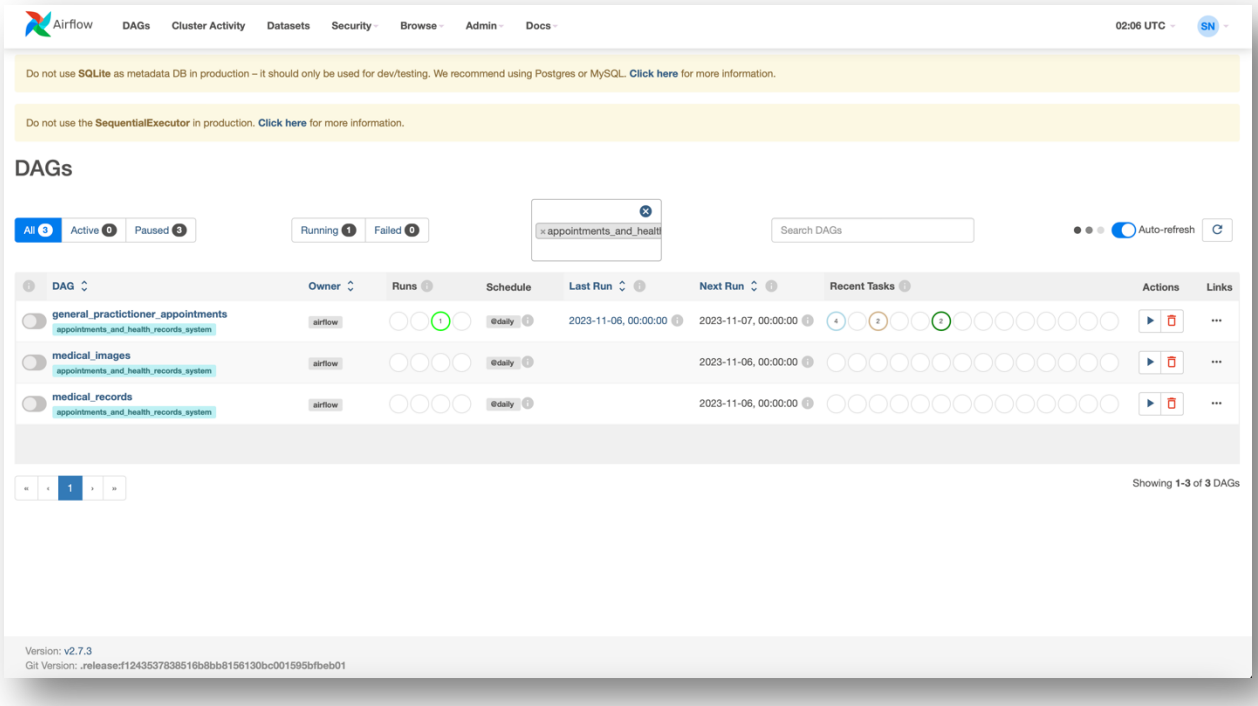
1. **Github Repository:** https://github.com/Sifael/big_data_challenge_costructor_university

The github repository link contains all technical assets and details for the implementation of the project.

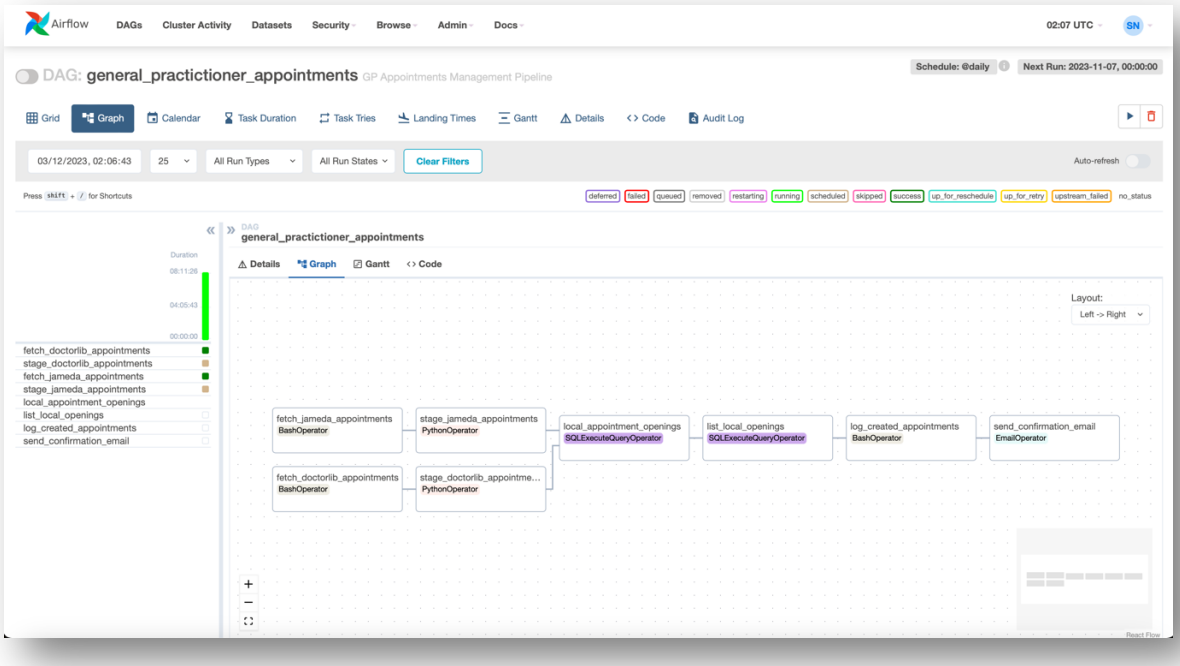
2. Data Sources and Description



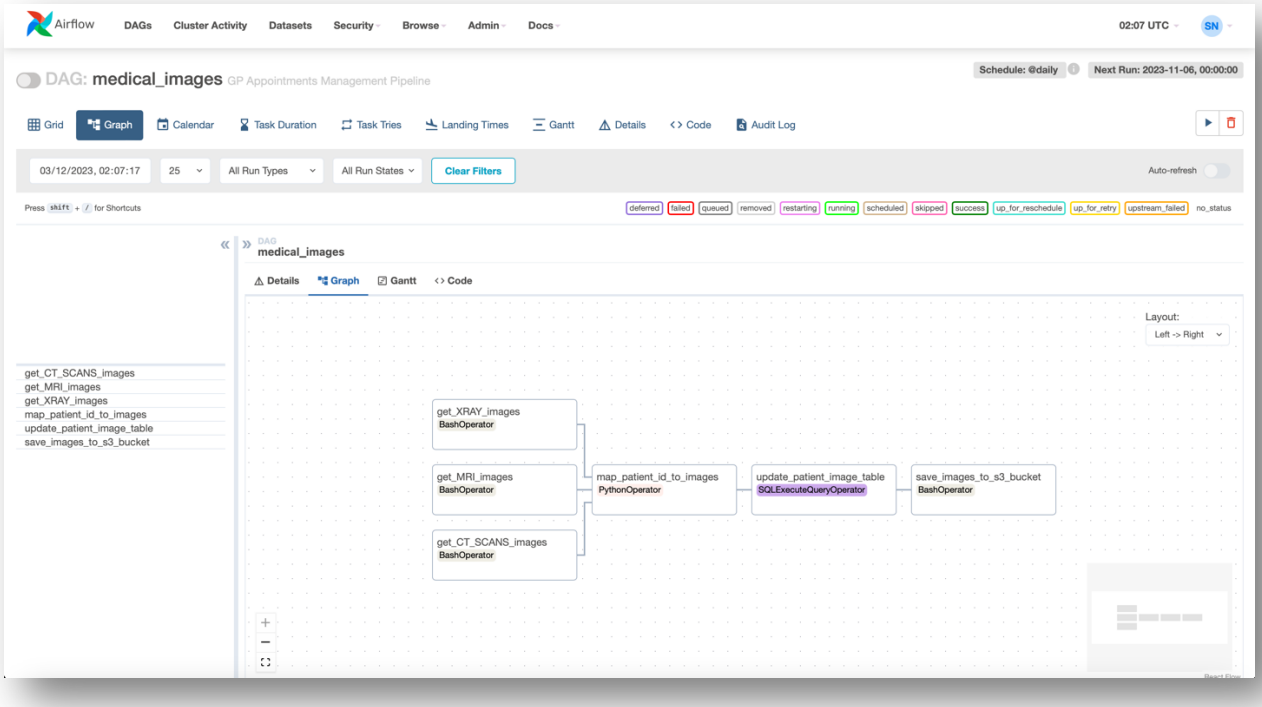
3. All Data Pipelines in Airflow DAG



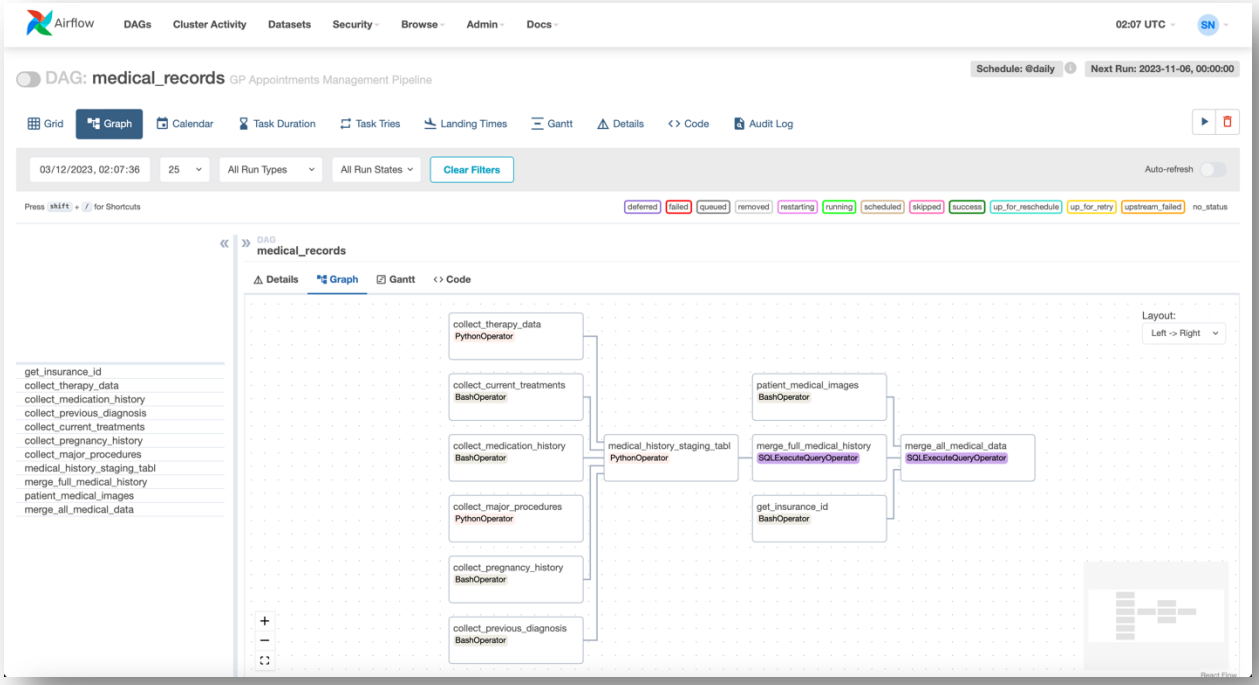
4. Appointment System Pipeline DAG



5. Medical Images Pipeline DAG



6. Medical Records Pipeline DAG



7. Predictive Analytics Pipeline DAG